

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from google.colab import drive
drive.mount('/content/drive')
```

*# Adjust the path to where your CSV is located*

```
df =
pd.read_csv('/content/drive/MyDrive/Cars_edmund_reviews_data.csv')
```

```
df.head()
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

```
{"summary":{"\n  \"name\": \"df\",\n  \"rows\": 40071,\n  \"fields\": [\n    {\n      \"column\": \"Unnamed: 0.1\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 12278,\n        \"min\": 0,\n        \"max\": 42287,\n        \"num_unique_values\": 40071,\n        \"samples\": [\n          22635,\n          26405,\n          34858\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"Unnamed: 0\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 12278,\n        \"min\": 0,\n        \"max\": 42287,\n        \"num_unique_values\": 40071,\n        \"samples\": [\n          22635,\n          26405,\n          34858\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"date\",\n      \"properties\": {\n        \"dtype\": \"object\",\n        \"num_unique_values\": 1345,\n        \"samples\": [\n          \"11/18/2006\",\n          \"09/11/2009\",\n          \"12/12/2008\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"author\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 27914,\n        \"samples\": [\n          \"Joe C\",\n          \"Jeep fan\",\n          \"Tundyguy\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"text\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 39310,\n        \"samples\": [\n          \"I recently purchased a bare bones work truck for a third vehicle. I have owned about every diffent model of truck in the last 20 years and I have to say I am very pleased with this truck. Dodge has managed to build a simple, very well built work truck that at an acceptional value. The six- speed manual allows for optimum power out of the six cylinder motor and provides good economy. I have averaged approximately 19 mpg round about and over 23 mpg on the

```

highway. To describe this vehicle in a few short words; good price, quiet at high speeds, plenty of power, great fuel economy for a full size truck. The standard cab has plenty of usable space behind the seat. The six speed manual is a great feature that allows for improved fuel economy. Plenty of cup holders and large door pockets allow for additional storage space. \", \n

\"A great car overall. Very powerful 3.5L engine. Drives great except in snow it's not very good. Lots of wheel spin, etc. Interior is well designed except no passenger power seat for \$30k?? Strange. Good gas mileage at 24 in mixed highway/city driving. Push button and door sensors are wonderful... never take your keys out of your pocket. That really spoils you! Push button start. Power--it MOVES (3.5L). Look of interior and exterior. Trunk space. \", \n

\"This our first BMW and probably will not be our last. First impression is very favorable. This car has great balance and performance and room for 4 and iDrive isn't the nightmare some make it out to be. Trunk size works for us. A large suitcase fits sideways under the top spacer along with my wife's makeup case and odds and ends. Plenty for a long weekend. My biggest gripe is with BMW making a decent iPod connection and sat radio an option. On a car in this price range in today's world, these should be standard. The wheels are ok, almost \"\"boring\"\". Fit and finish is good, just one dash rattle. Engine, interior, top operation, stereo, overall performance with comfort. \", \n

\"semantic\_type\": \"\", \n

\"description\": \"\", \n

\"column\": \"Year\", \n

\"properties\": { \n

\"dtype\": \"number\", \n

\"std\": 0, \n

\"min\": 2007, \n

\"max\": 2009, \n

\"num\_unique\_values\": 3, \n

\"samples\": [ \n

2007, \n

2008, \n

2009 \n

], \n

\"semantic\_type\": \"\", \n

\"description\": \"\", \n

} \n

{ \n

\"column\": \"Filename\", \n

\"properties\": { \n

\"dtype\": \"category\", \n

\"num\_unique\_values\": 596, \n

\"samples\": [ \n

\"2007\_subaru\_legacy.csv\", \n

\"2008\_audi\_tt.csv\", \n

\"2009\_bmw\_3\_series.csv\" \n

], \n

\"semantic\_type\": \"\", \n

\"description\": \"\", \n

} \n

{ \n

\"column\": \"cleaned\_text\", \n

\"properties\": { \n

\"dtype\": \"string\", \n

\"num\_unique\_values\": 39310, \n

\"samples\": [ \n

\"i recent purchas bare bone work truck third vehicl i own everi diffent model truck last 20 year i say i pleas truck dodg manag build simpl well built work truck accept valu the six speed manual allow optimum power six cylind motor provid good economi i averag approxim 19 mpg round 23 mpg highway to describ vehicl short word good price quiet high speed plenti power great fuel economi full size truck the standard cab plentli usabl space behind seat the six speed manual great featur allow improv fuel economi pleanti cup holder larg door pocket allow addit storag space \", \n

\"a great car overal veri power 3 5l engin drive great except snow not good lot wheel spin etc interior well design except no passeng power seat 30k strang good gas mileag 24 mix highway citi drive push button

```

door sensor wonder never take key pocket that realli spoil you push
button start power it move 3 5l look interior exterior trunk space",\
n      "\"this first bmw probabl not last first impress favor this
car great balanc perform room 4 idriv isn t nightmar make be trunk
size work us a larg suitcas fit sideways top spacer along wife s makeup
case odd end plenti long weekend my biggest gripe bmw make decent ipod
connect sat radio option on car price rang today s world standard the
wheel ok almost bore fit finish good one dash rattl engin interior top
oper stereo overal perform comfort\"\\n      ],\\n
\"semantic_type\\\": \\\"\\\",\\n      \"description\\\": \\\"\\\"\\n      }\\
n      },\\n      {\\n      \"column\\\": \\\"VADER_clean_text\\\",\\n
\"properties\\\": {\\n      \"dtype\\\": \\\"string\\\",\\n
\"num_unique_values\\\": 39310,\\n      \"samples\\\": [\\n      \"I
recently purchased a bare bones work truck for a third vehicle. I have
owned about every diffent model of truck in the last 20 years and I
have to say I am very pleased with this truck. Dodge has managed to
build a simple, very well built work truck that at an acceptional
value. The six- speed manual allows for optimum power out of the six
cylinder motor and provides good economy. I have averaged approximately
19 mpg round about and over 23 mpg on the highway. To describe this
vehicle in a few short words; good price, quiet at high speeds, plenty
of power, great fuel economy for a full size truck. The standard cab
has plenty of usable space behind the seat. The six speed manual is a
great feature that allows for improved fuel economy. Pleanty of cup
holders and large door pockets allow for additional storage space.\\\",\\
n      \"A great car overall. Very powerful 3.5L engine. Drives
great except in snow it's not very good. Lots of wheel spin, etc.
Interior is well designed except no passenger power seat for $30k??
Strange. Good gas mileage at 24 in mixed highway/city driving. Push
button and door sensors are wonderful... never take your keys out of
your pocket. That really spoils you! Push button start. Power--it
MOVES (3.5L). Look of interior and exterior. Trunk space.\\\",\\n
\"This our first BMW and probably will not be our last. First
impression is very favorable. This car has great balance and
performance and room for 4 and iDrive isn't the nightmare some make it
out to be. Trunk size works for us. A large suitcase fits sideways
under the top spacer along with my wife's makeup case and odds and
ends. Plenty for a long weekend. My biggest gripe is with BMW making a
decent iPod connection and sat radio an option. On a car in this price
range in today's world, these should be standard. The wheels are ok,
almost \\\"\\\"boring\\\"\\\". Fit and finish is good, just one dash rattle.
Engine, interior, top operation, stereo, overall performance with
comfort.\\\"\\n      ],\\n      \"semantic_type\\\": \\\"\\\",\\n
\"description\\\": \\\"\\\"\\n      }\\n      },\\n      {\\n      \"column\\\":
\"Vader_rating\\\",\\n      \"properties\\\": {\\n      \"dtype\\\":
\"number\\\",\\n      \"std\\\": 0.4110588666452841,\\n      \"min\\\": -
0.9907,\\n      \"max\\\": 0.9989,\\n      \"num_unique_values\\\":
6272,\\n      \"samples\\\": [\\n      0.2066,\\n      0.1724,\\n
-0.8225\\n      ],\\n      \"semantic_type\\\": \\\"\\\",\\n

```

```

{"description": "", "column": "VADER_sentiment", "properties": {"dtype": "number", "std": 0, "min": 0, "max": 1, "num_unique_values": 2, "samples": [0, 1]}, "semantic_type": "VADER_sentiment", "column": "affin_sentiment", "properties": {"dtype": "number", "std": 0, "min": 0, "max": 1, "num_unique_values": 2, "samples": [0, 1]}, "semantic_type": "affin_sentiment", "column": "TextBlob_sentiment", "properties": {"dtype": "number", "std": 0, "min": 0, "max": 1, "num_unique_values": 2, "samples": [0, 1]}, "semantic_type": "TextBlob_sentiment", "column": "final_sentiment", "properties": {"dtype": "number", "std": 0, "min": 0, "max": 1, "num_unique_values": 2, "samples": [0, 1]}, "semantic_type": "final_sentiment"}], "type": "dataframe", "variable_name": "df"}

print("2. Number of unique authors:", df['author'].nunique())
2. Number of unique authors: 27914

print("1. Total number of reviews:", len(df))
1. Total number of reviews: 40071

print("3. Most common review year:", df['Year'].mode()[0])
3. Most common review year: 2007

print("4. Average Vader rating:", df['Vader_rating'].mean())
4. Average Vader rating: 0.7913600658830575

print("5. Sentiment counts:\n", df['final_sentiment'].value_counts())
5. Sentiment counts:
final_sentiment
1    37885
0     2186
Name: count, dtype: int64

highest_vader_review = df.loc[df['Vader_rating'].idxmax()]
print("6. Review with highest Vader rating:\n",
highest_vader_review[['author', 'text', 'Vader_rating']])

```

```

6. Review with highest Vader rating:
  author                                D in Arkansas
  text                                I wanted a truck from another manufacture, but...
  Vader_rating                        0.9989
  Name: 34046, dtype: object

lowest_vader_review = df.loc[df['Vader_rating'].idxmin()]
print("7. Review with lowest Vader rating:\n",
lowest_vader_review[['author', 'text', 'Vader_rating']])

7. Review with lowest Vader rating:
  author                                Linda
  text                                I own a 2008 Kia Sportage EX which my daughter...
  Vader_rating                        -0.9907
  Name: 25523, dtype: object

print("8. Number of reviews per year:\n", df['Year'].value_counts())

8. Number of reviews per year:
  Year
2007    18048
2008    14359
2009     7664
  Name: count, dtype: int64

print("9. Top 5 authors:\n", df['author'].value_counts().head(5))

9. Top 5 authors:
  author
Mike      278
John      221
Dave      215
Chris     151
Steve     144
  Name: count, dtype: int64

print("10. Missing values:\n", df.isnull().sum())

10. Missing values:
  Unnamed: 0.1      0
  Unnamed: 0        0
  date            0
  author          2
  text            0
  Year            0
  Filename         0
  cleaned_text     0
  VADER_clean_text 0
  Vader_rating     0
  VADER_sentiment  0
  affin_sentiment  0

```

```

TextBlob_sentiment    0
final_sentiment        0
dtype: int64

diff_affin_vader = (df['affin_sentiment'] !=
df['VADER_sentiment']).sum()
print("11. Reviews with different affin and VADER sentiments:",
diff_affin_vader)

11. Reviews with different affin and VADER sentiments: 2621

print("12. Average Vader rating per year:\n", df.groupby('Year')
['Vader_rating'].mean())

12. Average Vader rating per year:
Year
2007    0.776572
2008    0.801164
2009    0.807817
Name: Vader_rating, dtype: float64

highest_avg_year = df.groupby('Year')['Vader_rating'].mean().idxmax()
print("13. Year with highest average Vader rating:", highest_avg_year)

13. Year with highest average Vader rating: 2009

diff_textblob_final = (df['TextBlob_sentiment'] !=
df['final_sentiment']).sum()
print("14. Reviews where TextBlob and final sentiment differ:",
diff_textblob_final)

14. Reviews where TextBlob and final sentiment differ: 1657

duplicate_cleaned_text = df.duplicated('cleaned_text').sum()
print("15. Number of duplicate cleaned_text entries:",
duplicate_cleaned_text)

15. Number of duplicate cleaned_text entries: 761

print("16. Number of reviews before 2010:", (df['Year'] < 2010).sum())

16. Number of reviews before 2010: 40071

print("17. Vader_rating Statistics:")
print("Mean:", df['Vader_rating'].mean())
print("Median:", df['Vader_rating'].median())
print("Standard Deviation:", df['Vader_rating'].std())

17. Vader_rating Statistics:
Mean: 0.7913600658830575

```

Median: 0.9509  
Standard Deviation: 0.4110588666452841

```
df['date'] = pd.to_datetime(df['date'], errors='coerce') # convert date column
```

```
earliest_reviews = df.sort_values('date').head(10)
```

```
print("18. 10 earliest reviews:\n", earliest_reviews[['author', 'date', 'text']])
```

18. 10 earliest reviews:

	author	date	\
2914	bob lucky	2006-01-18	
2913	Jason Smith	2006-01-23	
2912	07 Tahoe Owner	2006-01-24	
2911	Angler70	2006-01-26	
2910	Steve Smith	2006-01-31	
2909	Jacki	2006-02-03	
2908	Luke Hansen	2006-02-04	
2906	John	2006-02-05	
2907	Joe	2006-02-05	
2905	slightbo	2006-02-06	

	text
2914	This is a great driving vehicle. The performan...
2913	GMC has redeemed themselves with this outstand...
2912	The best thing about this vehicle is when you ...
2911	3rd row seats are useless unless you are a mid...
2910	I just bought my new Tahoe, and it is amazing....
2909	I never thought a vehicle could be so comforta...
2908	Incredible, that explains it! The vehicle is b...
2906	LTZ looks and drives great. Have driven Tahoes...
2907	I've been waiting for this one for awhile, and...
2905	Granted we have only had the car for a little ...

```
sentiment_distribution = df.groupby('Year')
```

```
['final_sentiment'].value_counts().unstack(fill_value=0)
```

```
print("19. Sentiment distribution by year:\n", sentiment_distribution)
```

19. Sentiment distribution by year:

final_sentiment	0	1
Year		
2007	1081	16967
2008	752	13607
2009	353	7311

```
years_negative_more =
```

```
sentiment_distribution[sentiment_distribution.get(1, 0) <  
sentiment_distribution.get(-1, 0)]
```

```
print("20. Years where negative reviews > positive reviews:\n",  
years_negative_more.index.tolist())
```

```
20. Years where negative reviews > positive reviews:  
[]
```